CS/CNS/EE/IDS 165: Foundations of Machine Learning and Statistical Inference

Introduction to Detection

http://tensorlab.cms.caltech.edu/users/anima/cms165-2020.html

Anima Anandkumar
Computing and Mathematical Sciences
California Institute of Technology, Pasadena CA 91125
anima@caltech.edu

Outline

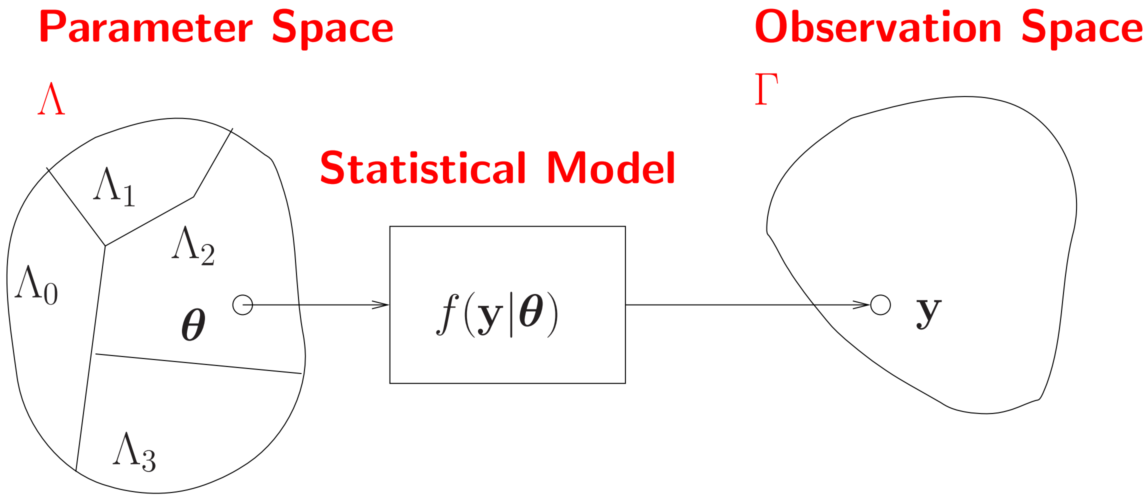
Concepts

- Detection and hypothesis testing.
- Deterministic and randomized detectors.
- Cost and risk.
- Bayesian, minimax, and Neyman-Pearson (NP) detectors.

References

1. H.V. Poor, An Introduction to Signal Detection and Estimation, 2nd Ed., Springer-Verlag, 1994, Chapter II.

The Hypothesis Testing Model



The Parameter Space: Λ The parameter $\boldsymbol{\theta}$ is chosen from one of the K disjoint subsets

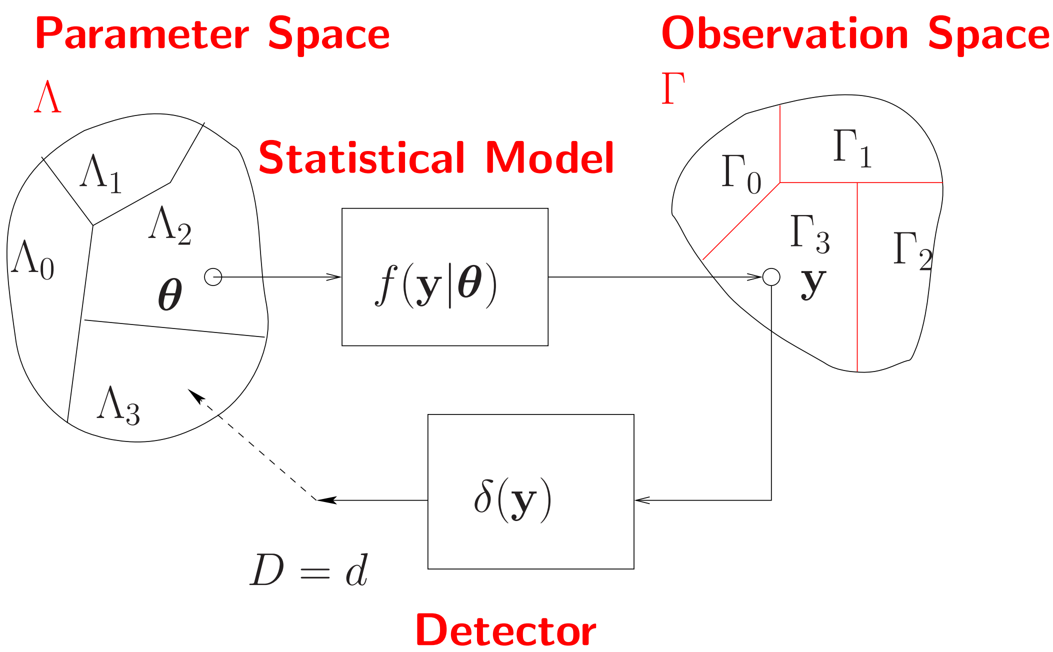
$$\Lambda = \Lambda_0 \bigcup \cdots \bigcup \Lambda_{K-1}, \quad \Lambda_i \bigcap \Lambda_j = \emptyset$$

Each Λ_i represents hypothesis $\mathcal{H}_i$. If Λ_i contains a single element θ_i, the hypothesis is simple. Otherwise, it is composite. When the hypotheses are composite, the parameter may be random with (conditional) distribution $\pi_i(\boldsymbol{\theta})$ within each distribution. If in addition we know the prior probability of each hypothesis, we have

$$\boldsymbol{\Theta} \sim \sum_{i=0}^{K-1} \Pr(\boldsymbol{\theta} \in \Lambda_i)\pi_i(\boldsymbol{\theta})$$

The Statistical Model: $(\Gamma, \mathcal{F}, f(\mathbf{y}|\boldsymbol{\theta}))$
For fixed $\boldsymbol{\theta} \in \Lambda$, the observation vector $\mathbf{Y} = (Y_1, \cdots, Y_n)$ is distributed according to $\mathbf{Y} \sim f(\mathbf{y}|\boldsymbol{\theta})$.

The Detection Problem



The Detection (Hypothesis Testing) Problem: Given realization (data) $\mathbf{Y} = \mathbf{y}$, detect to which Λ_i parameter $\boldsymbol{\theta}$ belongs, or equivalently, test hypotheses $\mathcal{H}_i : \mathbf{Y} \sim f(\mathbf{y}|\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Lambda_i$.

Detector: A deterministic detector $\delta(\cdot)$ maps observation $\mathbf{y}$ to a hypothesis index in $\{0, \cdots, K-1\}$. It partitions the observation space Γ into K disjoint subsets Γ_i and identify Γ_i with hypothesis $\mathcal{H}_i$. When the hypotheses are simple, $\Lambda_i = \{\theta_i\}$, $\delta : \Lambda \to \{\theta_i\}$.

A randomized detector† $\delta(\mathbf{y})$ maps observation $\mathbf{y}$ to a probability distribution $\delta(\mathbf{y}) = [\delta_1(\mathbf{y}), \cdots, \delta_K(\mathbf{y})]$ on $\{0, \cdots, K-1\}$ where $D \sim \delta(\mathbf{y})$ and

$$\delta_i(\mathbf{y}) = \Pr(D = i|\mathbf{Y} = \mathbf{y})$$

†Deterministic detectors are special cases of randomized detectors

Example: Detecting Signals Under Gaussian Noise

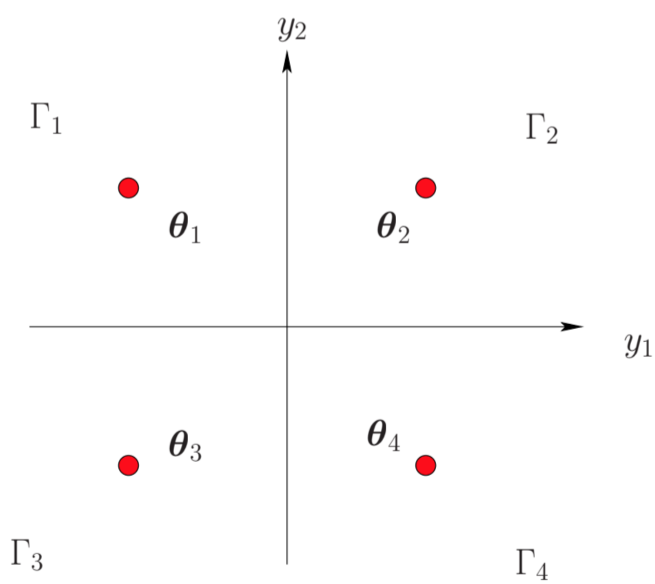
- The possible signals: $\Lambda = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4\} \in \mathcal{R}^2$.
- The observation: $\mathbf{Y} = \boldsymbol{\theta} + \mathbf{N}, \quad \mathbf{N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- The likelihood function: $f(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{2\pi} \exp\{-\frac{||\mathbf{y}-\boldsymbol{\theta}||^2}{2}\}$
- The ML detector is given by

$$\delta(\mathbf{y}) = \arg\max_{\boldsymbol{\theta}\in\Lambda} f(\mathbf{y}|\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}\in\Lambda} ||\boldsymbol{\theta} - \mathbf{y}||$$

An alternative description is the partition of $\mathcal{R}^2$ by

$$\Gamma_i = \{\mathbf{y} : ||\mathbf{y} - \theta_i|| < ||\mathbf{y} - \theta_j||, \forall j \neq i.\}$$

The points on the boundary are assigned arbitrarily.



Example: Detecting the Bias of a Coin

- The Parameter Space: Λ

$$\Lambda = \{\theta_0\} \bigcup \{\theta_1\}$$

where $\theta_0 =$ "HT", $\theta_1 =$ "HH".

- The Observation space $\Gamma = \{\mathtt{H}, \mathtt{T}\}$
- The statistical model:

$$f(y|\theta_0) = \begin{cases} 0.5 & y = \mathtt{H} \\ 0.5 & y = \mathtt{T} \end{cases} \qquad f(y|\theta_1) = \begin{cases} 1 & y = \mathtt{H} \\ 0 & y = \mathtt{T} \end{cases}$$

- A randomized strategy for detecting bias may be given by the probability $\delta(y)$ of the detection $D = 1$ ($\mathtt{HH}$), *i.e.,*

$$\delta(y) \triangleq \Pr(D = 1|Y = y)$$

Consider, for example,

$$\delta(y) = \begin{cases} 0 & y = \mathtt{T} \\ p & y = \mathtt{H} \end{cases}$$

where $p \in [0, 1]$ is the probability that θ_1 is chosen when $y = \mathtt{H}$.

- The detection is given by the Bernoulli trial $D \sim \mathcal{B}(\delta(y))$.

Cost and Risk

Cost: Each detection $D = d$ is associated with a cost $C(d, \boldsymbol{\theta})$. Since D is random, the cost $C_\delta(D, \boldsymbol{\theta})$ is random variable, and it is a function of $\boldsymbol{\theta}$.

Risk: The risk $R_{\boldsymbol{\theta}}(\delta)$ of a detector δ is the average cost (over the observation and the randomized variables). For a deterministic detector

$$R_{\boldsymbol{\theta}}(\delta) = \mathbb{E}(C(D, \boldsymbol{\theta})) = \int f(\mathbf{y}|\boldsymbol{\theta}) C(\delta(\mathbf{y}), \boldsymbol{\theta}) d\mathbf{y}$$

For a randomized detector, the detector is defined by the PMF $\delta(\mathbf{y}) = [\Pr(D = 0|\mathbf{Y} = \mathbf{y}), \cdots, \Pr(D = K-1|\mathbf{Y} = \mathbf{y})]$, and the risk is

$$R_{\boldsymbol{\theta}}(\delta) = \mathbb{E}(C(D, \boldsymbol{\theta})) = \int f(\mathbf{y}|\boldsymbol{\theta}) \sum_k \underbrace{\Pr(D = k|\mathbf{Y} = \mathbf{y})}_{\delta_k(\mathbf{y})} C_\delta(k, \boldsymbol{\theta}) d\mathbf{y}$$

Example: Consider $\Theta = \{\theta_0, \theta_1\}$. Define the cost by

	$\theta = \theta_0$	$\theta = \theta_1$
$D = \theta_0$	0	1
$D = \theta_1$	1	0

The risk when $\theta = \theta_0$ is given by

$$R_{\theta_0}(\delta) = \mathbb{E}(C(D, \theta_0)) = \int_{\{\mathbf{y}:\delta(\mathbf{y}) \neq 0\}} f(\mathbf{y}|\theta_0) d\mathbf{y} = \Pr(\delta(\mathbf{Y}) \neq \theta_0)$$

which is the probability of detection error (when $\theta = \theta_0$).

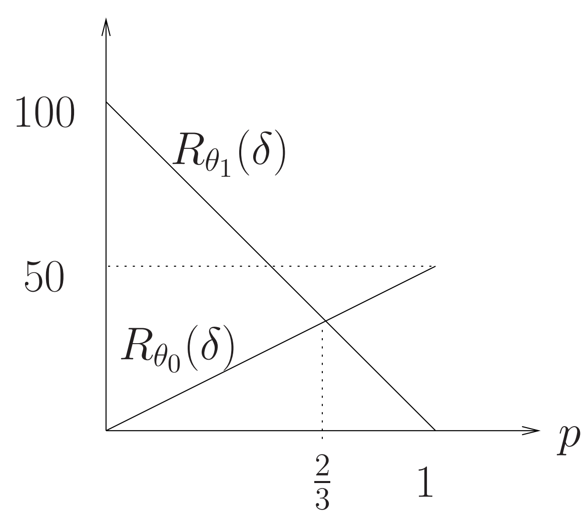
Cost and Risk: Example

Example: For the bias detection problem, the cost $C(\hat{\theta}, \theta)$ of detection $D \sim \delta(y)$ is given by

	$\theta = \theta_0$	$\theta = \theta_1$
$D = \theta_0$	0	100
$D = \theta_1$	100	0

For the randomized detector $\hat{\theta} \sim \mathcal{B}(\delta(y))$ defined previously,

$$\begin{aligned} R_{\theta_0}(\delta) &= \Pr(y=T;\theta_0)[C(\theta_0,\theta_0)\Pr(D=\theta_0|y=T) + C(\theta_1,\theta_0)\Pr(D=\theta_1|y=T)] \\ &\quad + \Pr(y=H;\theta_0)[C(\theta_0,\theta_0)\Pr(D=\theta_0|y=H) + C(\theta_1,\theta_0)\Pr(D=\theta_1|y=H)] \\ &= 50p \\ R_{\theta_1}(\delta) &= \Pr(y=T;\theta_1)[C(\theta_0,\theta_1)\Pr(D=\theta_0|y=T) + C(\theta_1,\theta_1)\Pr(D=\theta_1|y=T)] \\ &\quad + \Pr(y=H;\theta_1)[C(\theta_0,\theta_1)\Pr(D=\theta_0|y=H) + C(\theta_1,\theta_1)\Pr(D=\theta_1|y=H)] \\ &= 100(1-p) \end{aligned}$$



The Bayesian Detector

Bayesian Risk and Bayesian Detector

The Bayesian formulation assumes a prior distribution of the parameter $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. The Bayesian risk is the risk averaged over priors:

$$R(\delta) = \int \pi(\boldsymbol{\theta}) R_{\boldsymbol{\theta}}(\delta) d\boldsymbol{\theta}$$

Note that the Bayesian risk no longer depends on the parameter. The Bayesian detector minimizes the Bayesian Risk

$$\min_{\delta} \mathbb{E}(R_{\boldsymbol{\theta}}(\delta)) = \min_{\delta} \int \pi(\boldsymbol{\theta}) R_{\boldsymbol{\theta}}(\delta) d\boldsymbol{\theta}$$

Example:

Let $\Theta = \{\theta_0, \theta_1\} \sim \pi(\theta)$. Define the cost by

$$C(\delta(\mathbf{y}), \theta_i) = \begin{cases} 1 & \delta(\mathbf{y}) = i \\ 0 & o.w. \end{cases}$$

The risk $R_\theta(\delta)$ is the conditional error probability

$$R_\theta(\delta) = E_\theta(C(\delta(\mathbf{Y}), \theta)) = \int_{\{\mathbf{y}:\delta(\mathbf{y})\neq\theta\}} f(\mathbf{y}|\theta) d\mathbf{y} = \Pr(\delta(\mathbf{Y}) \neq \theta | \Theta = \theta)$$

The Bayesian Risk is the (unconditional) error probability

$$\begin{aligned} R(\delta) &= \pi(\theta_0) \Pr(\delta(\mathbf{Y}) \neq \theta_0 | \Theta = \theta_0) + \pi(\theta_1) \Pr(\delta(\mathbf{Y}) \neq \theta_1 | \Theta = \theta_1) \\ &= \Pr(\delta(\mathbf{Y}) \neq \Theta) \end{aligned}$$

Minimizing the Bayesian risk minimizes the detection error probability

The Minimax Detector

The Criterion

Minimize the risk for the worst case:

$$\min_{\delta} \max_{\boldsymbol{\theta} \in \Theta} R_{\boldsymbol{\theta}}(\delta)$$

Example:

For the problem of detecting the bias, the minimax detector is given by

$$\min_{\delta(y)} \max\{R_{\theta_0}(\delta), R_{\theta_1}(\delta)\}$$

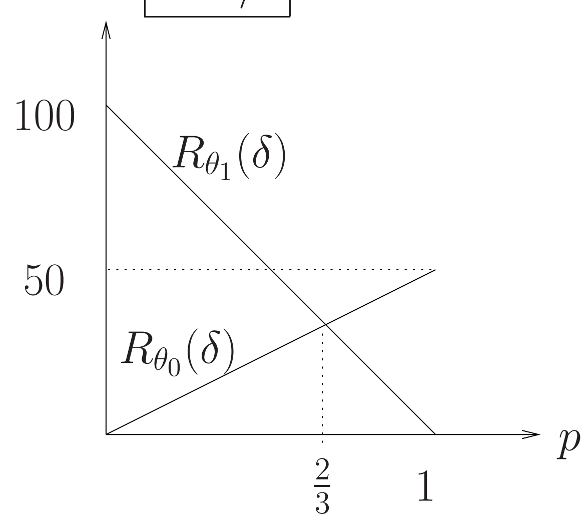
This is equivalent to

$$\min_{p} \max\{50p, 100(1-p)\}$$

The optimal p can be computed from

$$50p_{\mathsf{opt}} = 100(1-p_{\mathsf{opt}}) \to p_{\mathsf{opt}} = \frac{2}{3}$$

The maximum risk is $\boxed{\$100/3}$.



The Neyman-Pearson Detector

The Criterion

Minimize the risk for one parameter while imposing constraints on the risks for other parameters.

$$\min_{\delta} R_{\theta_K}(\delta) \text{ subject to } R_{\theta_i}(\delta) \leq \alpha_i, i < K$$

Example

For the problem of flipping a coin with an unknown bias, the Neyman-Pearson detector can be formulated as

$$\min_{\delta} R_{\theta_1}(\delta) \text{ subject to } R_{\theta_0}(\delta) \leq 10.$$

The optimal δ is given by $p = 1/5$.

A more common formulation is to use probabilities as risks.

$$\min_{\delta} \Pr[D \neq \theta_1; \theta_1] \text{ subject to } \Pr[\hat{\theta} \neq \theta_0; \theta_0] \leq \alpha$$

Outline

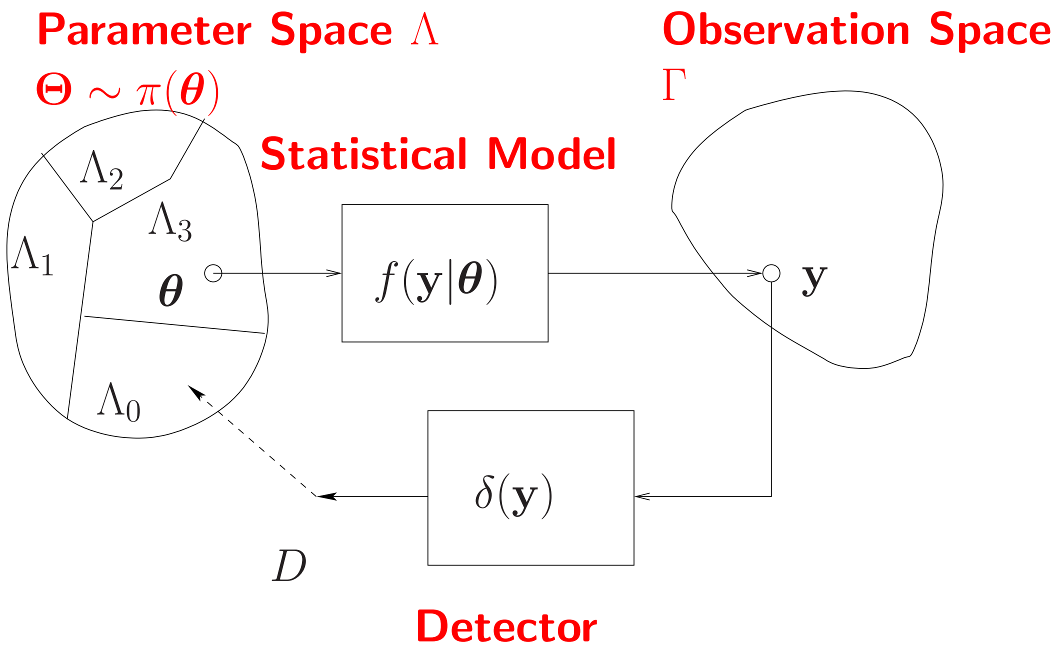
Main Concepts

- The Bayesian Formulation
- The Bayesian Risk and Conditional Risk.
- Testing Binary Hypothesis.
- Examples

References

1. H.V. Poor, An Introduction to Signal Detection and Estimation, 2nd Ed., Springer-Verlag, 1994, Chapter II.
2. S. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993, Chapters 3, 6.
3. L. L. Scharf, Statistical Signal Processing: Detection, Estimation and Time Series Analysis, Addison-Wesley, 1991, Chapter 5.
4. H.L. Van Trees, Detection, Estimation, and Modulation Theory, vol. I. Wiley, New York, 1968, Chap. 2.
5. E.L. Lehmann, Testing Statistical Hypothesis, Wiley, 1986, Chapter 1.

The Bayesian Formulation



The Prior and the Statistical Model

We assume that the random parameter $\boldsymbol{\Theta}$ has a prior distribution $\boldsymbol{\Theta} \sim \pi(\boldsymbol{\theta})$, which leads to

$$\pi_i \stackrel{\Delta}{=} \Pr(\boldsymbol{\Theta} \in \Lambda_i) = \int_{\Lambda_i} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

The statistical model is defined by

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Recall the problem statement: Given $\mathbf{Y} = \mathbf{y}$, test K-ary hypothesis

$$\mathcal{H}_i : \mathbf{Y} \sim f(\mathbf{y}; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Lambda_i, \quad i = 0, \cdots, K-1$$

The (randomized) detector is defined by

$$\boldsymbol{\delta}(\mathbf{y}) = [\delta_k(\mathbf{y})], \quad \delta_k(\mathbf{y}) \stackrel{\Delta}{=} \Pr(D = k|\mathbf{Y} = \mathbf{y}).$$

From Prior to Posterior Distributions

Testing based on Priors

Suppose that we don't take any measurements. Our best bet would be based on the priors

$$\delta = \arg\max \pi_i, \qquad \pi_i = \int_{\Lambda_i} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Testing based on Posterior Distributions

Once we have $\mathbf{Y} = \mathbf{y}$, the odds have changed. The posterior distribution of $\boldsymbol{\theta}$ and posterior probabilities of the hypotheses are given by

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y})}, \quad \Pr(\boldsymbol{\theta} \in \Lambda_i|\mathbf{y}) = \frac{1}{f(\mathbf{y})} \int_{\Lambda_i} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Thus the detector that maximizes the probability of correct detection is given by the maximum a posteriori (MAP) detector:

$$\delta(\mathbf{y}) = \arg\max_i \int_{\Lambda_i} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$$

For $\Lambda_i = \{\boldsymbol{\theta}_i\}$ with prior $\pi_i = \Pr(\boldsymbol{\Theta} = \theta_i)$,

$$\delta(\mathbf{y}) = \arg\max_i \pi_i f(\mathbf{y}|\boldsymbol{\theta}_i)$$

For simple binary hypotheses, we have the likelihood ratio detector

$$\delta(\mathbf{y}) = \begin{cases} 1 & \frac{f(\mathbf{y}|\boldsymbol{\theta}_1)}{f(\mathbf{y}|\boldsymbol{\theta}_0)} > \frac{\pi_0}{\pi_1} \\ 0 & \text{otherwise} \end{cases}$$

Cost, Risk, and Bayesian Detector

Cost Function. Given $\boldsymbol{\theta}$, we define the cost function for each detection $D = i$ by $C(i, \boldsymbol{\theta})$. The cost is uniform if, for all $\boldsymbol{\theta} \in \Lambda_k$, $C(i, \boldsymbol{\theta}) = C_{ik}$, and the cost function is represented by matrix $\mathbf{C} = [C_{ij}]$.

The Risks. For each fixed parameter $\boldsymbol{\theta} \in \Lambda$, the risk of a detector δ is defined as

$$\begin{aligned} R_{\boldsymbol{\theta}}(\delta) &= \mathbb{E}(C(D, \boldsymbol{\theta}) | \boldsymbol{\Theta} = \boldsymbol{\theta}) = \sum_k \Pr(D = k) C(k; \boldsymbol{\theta}) \\ &= \sum_k C(k, \boldsymbol{\theta}) \int \delta_k(\mathbf{y}) f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y} \end{aligned}$$

The Bayesian risk is the risk averaged over prior $\pi(\boldsymbol{\theta})$

$$\begin{aligned} R(\delta) &= \mathbb{E}(R_{\boldsymbol{\Theta}}(\delta)) = \mathbb{E}(C(D, \boldsymbol{\Theta})) \\ &= \int f(\mathbf{y}) \mathbb{E}(C(D, \boldsymbol{\Theta}) | \mathbf{Y} = \mathbf{y}) d\mathbf{y} = \int f(\mathbf{y}) R(\delta | \mathbf{y}) d\mathbf{y} \end{aligned}$$

where $R(\delta|\mathbf{y}) \triangleq \mathbb{E}(C(\delta(\mathbf{y}), \boldsymbol{\Theta}) | \mathbf{Y} = \mathbf{y})$ is the conditional risk, which is a function of $\mathbf{y}$.

The Bayesian Detector The Bayesian detector minimizes the Bayesian risk

$$\min_\delta R(\delta) \leftrightarrow \min_\delta R(\delta|\mathbf{y}) \leftrightarrow \min_\delta \mathbb{E}(C(D, \boldsymbol{\Theta}) | \mathbf{Y} = \mathbf{y})$$

where we note that $D \sim \delta(\mathbf{y})$.

Some Heuristics

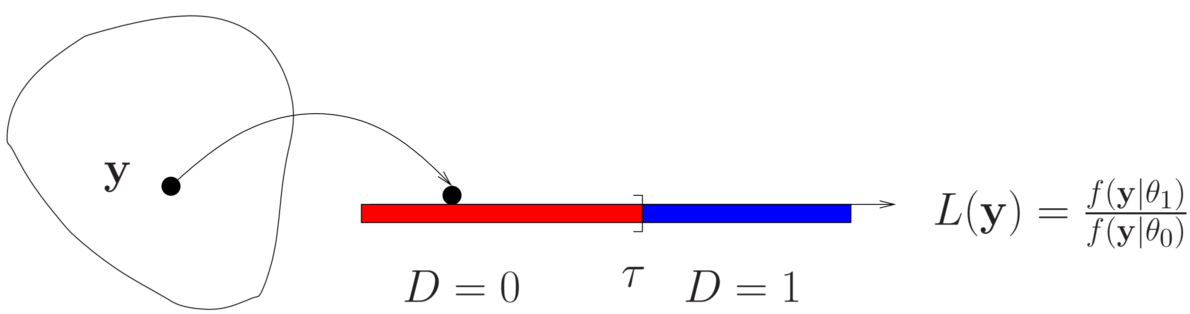
Consider simple binary hypotheses with $\Theta = \{\theta_0, \theta_1\}$ and with priors $\pi_i \triangleq \Pr(\Theta = \theta_i)$. We consider the following cost function

	θ_0	θ_1
$D = 0$	0	$C_{0,1}$
$D = 1$	$C_{1,0}$	0

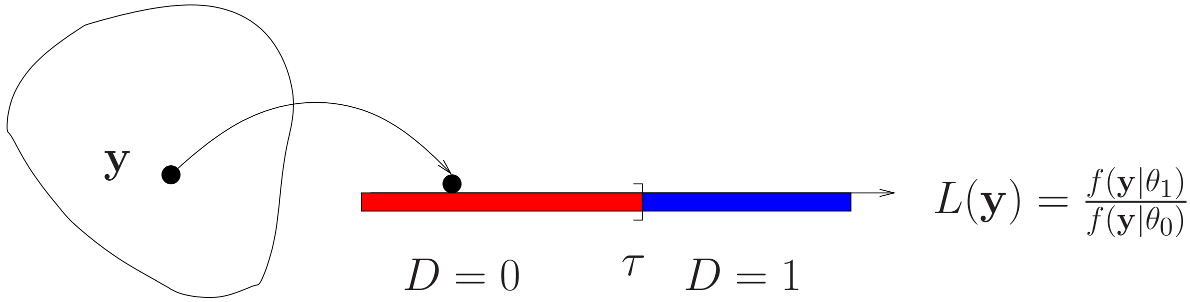
We now argue heuristically the form of the Bayesian detector.

- A sufficient statistic is the likelihood ratio
$$L(\mathbf{y}) \triangleq \frac{f(\mathbf{y}|\theta_1)}{f(\mathbf{y}|\theta_0)}$$
which maps the original n-D statistic to a 1-D statistic.

- It is intuitive that the larger the $L(\mathbf{y})$, the more we should favor θ_1. This suggests a threshold detector on $L(\mathbf{y})$ with the threshold τ depends on $\{\pi_i, C_{ij}\}$.



Some Heuristics



How to choose threshold τ?

- The threshold should be affected by the prior π_i. If $\pi_0 = \Pr(\Theta = \theta_0)$ is large, the region for detecting θ_0 should be larger, *i.e.,* τ larger. This should be made relative to π_1. One possible choice is

$$\tau = \frac{\pi_0}{\pi_1}$$

  In particular, such a choice makes sense when $\pi_0 = 0$.

- The threshold should also be affected by costs; if C_{01} is large, then the cost of mistakenly detecting θ_1 is high. Thus the region for detecting θ_1 should be reduced, and τ should be smaller. This again needs to be made relative to the cost C_{10}. One possible choice, incorporating the effect of priors, is

$$\tau = \frac{\pi_0}{\pi_1}\frac{C_{10}}{C_{01}}$$

  The choice above makes sense, but it gives no indication that it is in any way optimal.

The Optimal Bayesian Detector

Theorem

The Bayesian risk is minimized by a deterministic detector

$$\delta_k(\mathbf{y}) = \begin{cases} 1 & k = k_o \\ 0 & o.w. \end{cases} \tag{1}$$

where k_o minimizes the conditional cost

$$\begin{aligned} k_o &= \arg\min_k \mathbb{E}(C(k, \boldsymbol{\Theta})|\mathbf{Y} = \mathbf{y}) \\ &= \arg\min_k \int C(k, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \end{aligned}$$

Remark:

The Bayesian detector minimizes the cost averaged over parameter Θ and conditioned on data $\mathbf{y}$.

Proof: We fix $\mathbf{y}$ and we need to find the optimal

$$\delta_k(\mathbf{y}) \stackrel{\Delta}{=} \Pr(D = k|\mathbf{Y} = \mathbf{y}), \quad k = 0, \cdots, K-1$$

that minimizes

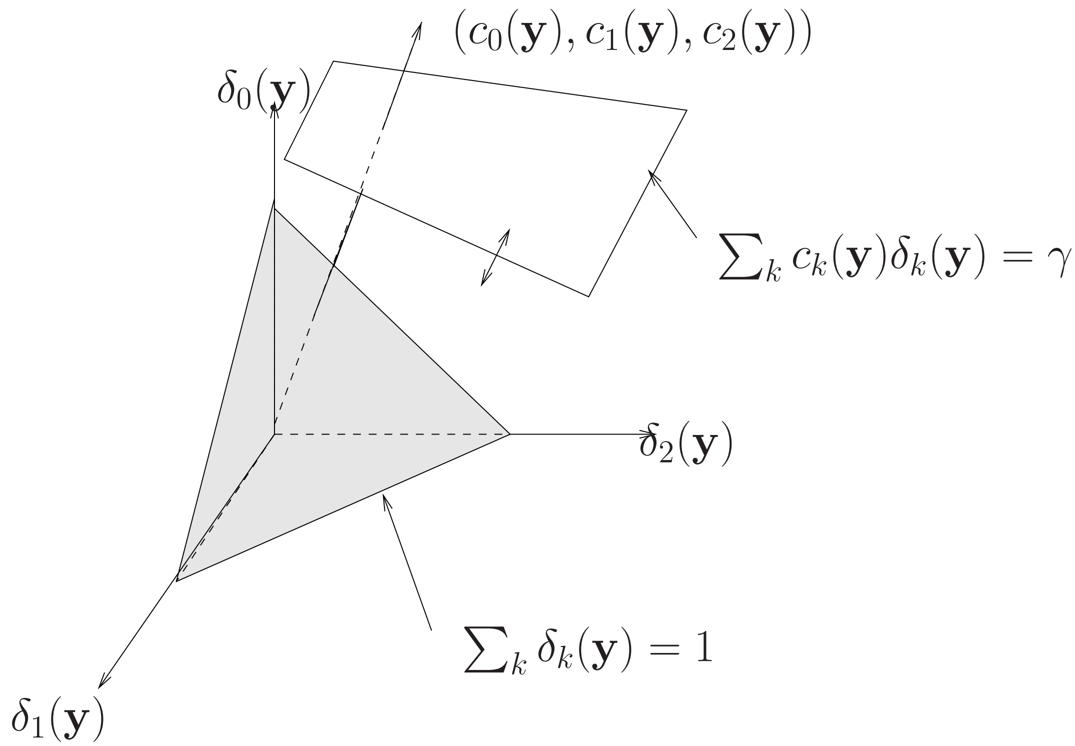
$$\begin{aligned} R(\delta|\mathbf{y}) &\stackrel{\Delta}{=} \mathbb{E}(C(D, \boldsymbol{\Theta})|\mathbf{Y} = \mathbf{y}) \\ &= \mathbb{E}(\mathbb{E}(C(D, \boldsymbol{\Theta})|D, \mathbf{Y} = \mathbf{y})) \\ &= \sum_k \Pr(D = k|\mathbf{Y} = \mathbf{y})\mathbb{E}(C(k, \boldsymbol{\Theta})|D = k, \mathbf{Y} = \mathbf{y})) \\ &= \sum_k \delta_k(\mathbf{y})\mathbb{E}(C(k, \boldsymbol{\Theta})|\mathbf{Y} = \mathbf{y})) = \sum_k \delta_k(\mathbf{y})c_k(\mathbf{y}) \end{aligned}$$

where the last equality comes from the fact that $\Theta \to \mathbf{Y} \to D$ forms a Markov chain, *i.e.*, conditioned on Y, Θ and D are independent.

We can formulate this optimization as a linear programming problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{k=0}^{K-1} \delta_k(\mathbf{y}) c_k(\mathbf{y}) \\ \text{subject to} \quad & \sum_{k=0}^{K-1} \delta_k(\mathbf{y}) = 1, \quad \delta_k(\mathbf{y}) \geq 0, \quad \forall k \end{aligned}$$

From the Fundamental Theorem of Linear Programming[†], The solution to this optimization is given in (1).



The linear programming problem can be illustrated geometrically. The minimization corresponds to moving the plane defined by $\sum_k c_k(\mathbf{y})\delta_k(\mathbf{y}) = \gamma$ up from $\gamma = 0$ until it touches first the the polyhedron defined by $\sum_k \delta_k(\mathbf{y}) = 1$. The optimal $\delta_k(\mathbf{y})$ must be one of the extreme points of the polyhedron.

[†] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd Ed., Addison-Wesley, 1989.

Simple Binary Hypothesis: Detector

For simple hypothesis $\Lambda = \{\theta_0, \theta_1\}$ with priors $\pi_i = \Pr(\theta = \theta_i)$ and cost function $C_{ij} \stackrel{\Delta}{=} C(i, \theta_j)$, the conditional costs are given by

$$
\begin{aligned}
\mathbb{E}(C(1,\theta)|\mathbf{Y} = \mathbf{y}) &= \frac{1}{f(\mathbf{y})}(C_{10} f(\mathbf{y}|\theta_0)\pi_0 + C_{11} f(\mathbf{y}|\theta_1)\pi_1) \\
\mathbb{E}(C(0,\theta)|\mathbf{Y} = \mathbf{y}) &= \frac{1}{f(\mathbf{y})}(C_{00} f(\mathbf{y}|\theta_0)\pi_0 + C_{01} f(\mathbf{y}|\theta_1)\pi_1)
\end{aligned}
$$

Hence $\delta(\mathbf{y}) = 1$ if

$$f(\mathbf{y}|\theta_1)\pi_1(C_{01} - C_{11}) \geq f(\mathbf{y}|\theta_0)\pi_0(C_{10} - C_{00})$$

Without loss of generality, assume $C_{ij} > C_{ii}$ for all i, j. We have

$$\delta(\mathbf{y}) = 1 \quad \text{if} \quad \frac{f(\mathbf{y}|\theta_1)}{f(\mathbf{y}|\theta_0)} \geq \frac{\pi_0(C_{10} - C_{00})}{\pi_1(C_{01} - C_{11})} \stackrel{\Delta}{=} \tau_B$$

Remarks

- The detector is given by the likelihood ratio (sufficient statistic for θ).
- If θ_0 is more likely, $\pi(\theta_0)$ is bigger and the threshold for detecting θ_1 is higher.
- When $C_{ii} = 0$, bigger the C_{10}, higher the threshold of detecting θ_1.
- When the two hypothesis are equally likely, and $C_{i1} = C_{i0}$, $\tau_B = 1$, which implies that the detector picks the hypothesis that is more likely.

Simple Binary Hypothesis: The Risk

The Decision Region

The optimal detector partitions the observation space into Γ_0 and Γ_1 according to

$$\Gamma_1 \stackrel{\Delta}{=} \{\mathbf{y} : f(y|\theta_1) \geq \tau_B f(y|\theta_0)\}, \Gamma_0 = \Gamma_1^c$$

The Bayesian Risk

The Bayesian risk of the optimal detector is given by

$$\begin{aligned} R(\delta) &= \pi_0[C_{00}\int_{\Gamma_0} f(\mathbf{y}|\theta_0)d\mathbf{y} + C_{10}\int_{\Gamma_1} f(\mathbf{y}|\theta_0)d\mathbf{y}] + \\ &\quad \pi_1[C_{01}\int_{\Gamma_0} f(\mathbf{y}|\theta_1)d\mathbf{y} + C_{11}\int_{\Gamma_0} f(\mathbf{y}|\theta_1)d\mathbf{y}] \end{aligned}$$

A Special Cost

For the 0-1 cost, *i.e.*, $C_{ij} = 1$ when $i \neq j$ and $C_{ii} = 0$, we have

$$\begin{aligned} R(\delta) &= \pi_0\int_{\Gamma_1} f(\mathbf{y}|\theta_0)d\mathbf{y} + \pi_1\int_{\Gamma_0} f(\mathbf{y}|\theta_1)d\mathbf{y} \\ &= \Pr(\text{detection error}) \end{aligned}$$

Example: Detection of 2 Signals

Signal in Gaussian Noise

Consider the transmission of one of the two signal θ_0 and θ_1 over a Gaussian channel

$$Y = \Theta + N, \Theta \in \{\theta_0, \theta_1\}, N \sim \mathcal{N}(0, \sigma^2).$$

The prior probability is given by $\pi_i = \Pr(\Theta = \theta_i)$.

The simple binary hypothesis can be expressed as

$$\begin{aligned} \mathcal{H}_0 : &\quad Y \sim \mathcal{N}(\theta_0, \sigma^2), \\ \mathcal{H}_1 : &\quad Y \sim \mathcal{N}(\theta_1, \sigma^2). \end{aligned}$$

Cost and Risk

Suppose that we have the cost, *i.e.,*

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

The Bayesian risk for this cost is the error probability.

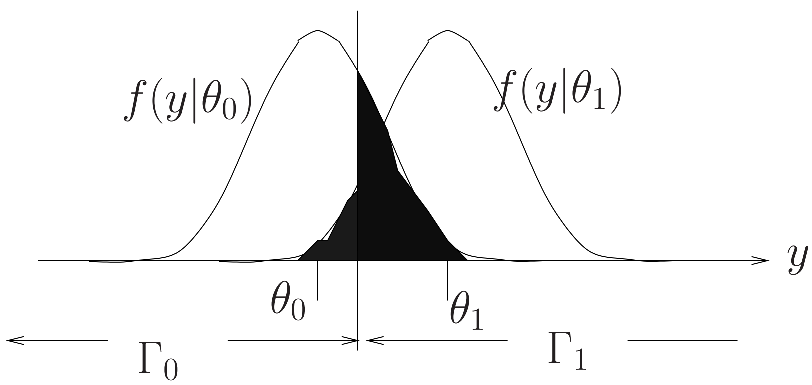
The Likelihood Functions

$$f(y|\theta_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(y-\theta_i)^2}{2\sigma^2}\}.$$

Decision Region

The observation space is partitioned into Γ_0 and Γ_1 where

$$\begin{aligned}
\Gamma_1 &\triangleq \{y : \frac{f(y|\theta_1)}{f(y|\theta_0)} \geq \frac{\pi_0}{\pi_1}\}. \\
&= \{y : \ln f(y|\theta_1) - \ln f(y|\theta_1) \geq \ln \frac{\pi_0}{\pi_1}\} \\
&= \{y : y \geq \frac{\theta_0 + \theta_1}{2} + \frac{\sigma^2}{\theta_1 - \theta_0} \ln \frac{\pi_0}{\pi_1} \triangleq \gamma\}
\end{aligned}$$



Remarks

- When $\pi_0 = \pi_1$, the decision boundary γ lies between the means of the two density functions. The boundary moves with π_i to favor the more likely hypothesis.
- For really noisy data, $\sigma \to \infty$, the detection goes with the prior. On the other hand, for really clean data, the detection ignores the prior.

Error Probability

The Baysian risk (the error probability) is given by

$$R(\delta) \;=\; \pi_0 Q(\frac{\gamma-\theta_0}{\sigma}) + \pi_1 Q(\frac{\theta_1-\gamma}{\sigma})$$

where the Q-function is the tail probability

$$Q(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

Remarks

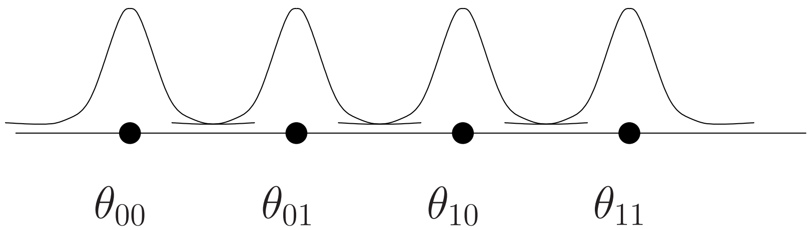
The error probability does not depend on the actual locations of θ_i; it depends on their relative distance $\theta_1 - \theta_0$:

$$\begin{aligned}
\frac{\gamma-\theta_0}{\sigma} &= \frac{\theta_1+\theta_0}{2\sigma} + \frac{\sigma}{\theta_1-\theta_0}\ln\frac{\pi_0}{\pi_1} \\
\frac{\theta_1-\gamma}{\sigma} &= \frac{\theta_1+\theta_0}{2\sigma} - \frac{\sigma}{\theta_1-\theta_0}\ln\frac{\pi_0}{\pi_1}
\end{aligned}$$

When the two hypotheses are equally likely, we have

$$\gamma = \frac{\theta_1+\theta_0}{2}, R(\delta) = Q(\frac{\theta_1-\theta_0}{2\sigma})$$

Example: Composite Binary Hypothesis



Consider the following binary hypothesis:

$$\mathcal{H}_0 : \quad Y = \Theta + N, \Theta \in \Lambda_0 = \{\theta_{00}, \theta_{01}\}$$
$$\mathcal{H}_1 : \quad y = \Theta + N, \Theta \in \Lambda_1 = \{\theta_{10}, \theta_{11}\}$$

where $N \sim \mathcal{N}(0, \sigma^2)$. Suppose also that $\pi_{ij} \triangleq \Pr(\Theta = \theta_{ij}) = \frac{1}{4}$, and the cost is uniform, *i.e.*,

$$C_{i,\theta} = \begin{cases} 0 & \theta \in \Lambda_i, \\ 1 & \text{otherwise} \end{cases}$$

The optimal detector is given by

$$\delta(y) = \begin{cases} 0 & \sum_{i,j} C(0, \theta_{ij}) f(\mathbf{y}|\theta_{ij}) \pi_{ij} < \sum_{i,j} C(1, \theta_{ij}) f(\mathbf{y}|\theta_{ij}) \pi_{ij} \\ 1 & \sum_{i,j} C(1, \theta_{ij}) f(\mathbf{y}|\theta_{ij}) \pi_{ij} < \sum_{i,j} C(0, \theta_{ij}) f(\mathbf{y}|\theta_{ij}) \pi_{ij} \end{cases}$$

Substituting π_{ij} and $C(i, \theta)$, the detector is given by the test the ratio

$$L(y) \;=\; \frac{\pi_{10} f(y|\theta_{10}) + \pi_{11} f(y|\theta_{11})}{\pi_{00} f(y|\theta_{00}) + \pi_{01} f(y|\theta_{01})} = \frac{e^{-\frac{(y-\theta_{10})^2}{2\sigma^2}} + e^{-\frac{(y-\theta_{11})^2}{2\sigma^2}}}{e^{-\frac{(y-\theta_{00})^2}{2\sigma^2}} + e^{-\frac{(y-\theta_{01})^2}{2\sigma^2}}}$$

Suppose that $\theta_{00} = -\theta_{11}$ and $\theta_{01} = -\theta_{10}$. Intuition suggests that

$$\Gamma_0 = \{y : y < 0\}, \quad \Gamma_1 = \{y : y \geq 0\},$$

and it is true.

Example: Detecting Bias

Recall the the coin tossing problem $Y \sim p(y;\theta)$ with binary hypotheses

$$\mathcal{H}_0 : f(y|\theta_0) = \begin{cases} 0.5 & y = \mathsf{H} \\ 0.5 & y = \mathsf{T} \end{cases} \quad \text{vs.} \mathcal{H}_1 : p(y;\theta_1) = \begin{cases} 1 & y = \mathsf{H} \\ 0 & y = \mathsf{T} \end{cases}$$

The cost is given by $C_{ii} \stackrel{\Delta}{=} C(i,\theta_i) = 0$ and $C_{ij} \stackrel{\Delta}{=} C(i,\theta_j) = 100$ for $j \neq i$. The Bayesian detector is based on testing the likelihood ratio against the threshold

$$\tau_B = \frac{\pi_0(C_{10} - C_{00})}{\pi_1(C_{01} - C_{11})} = \frac{\pi_0}{\pi_1}.$$

Because we have discrete random variables, we need to do this test individually for $y = T$ and $y = H$.

$$\begin{array}{ll} y = \mathsf{T} & \dfrac{f(\mathsf{T}|\theta_1)}{f(\mathsf{T}|\theta_0)} = 0 \leq \tau_B \to \delta(\mathsf{T}) = 0. \\ y = \mathsf{H} & \dfrac{f(\mathsf{H}|\theta_1)}{f(\mathsf{H}|\theta_0)} = 2 \stackrel{?}{>} \tau_B \end{array}$$

As expected, the Bayesian detector depends on the prior π_0.

$$\delta_B(\mathsf{T}) = 0, \quad \delta_B(\mathsf{H}) = \begin{cases} 0 & \pi_0 \geq \frac{2}{3} \\ 1 & \pi_0 < \frac{2}{3} \end{cases}$$

The optimal Bayesian detector is given by

$$\begin{aligned} R(\delta_B) &= E\{\delta(Y)C_{1,\theta} + (1 - \delta(Y))C_{0,\theta}\} \\ &= \pi_0 P(\delta_B = 1|\theta_0) \times 100 + \pi_1 P(\delta_B = 0|\theta_1) \times 100 \\ &= \begin{cases} 100\pi_1 & \pi_0 \geq \frac{2}{3} \\ 50\pi_0 & \pi_0 < \frac{2}{3} \end{cases} \to \max_{\pi_0} R(\delta_B) = 100/3 \end{aligned}$$

Summary

Assumption: $\mathbf{\Theta} \sim \pi(\boldsymbol{\theta})$

Key Steps: Minimizing the conditional cost:

$$d = \arg\min_i \mathbb{E}(C(i, \boldsymbol{\theta})|\mathbf{y})$$

where the average is taken over Θ using conditional distribution $f(\theta|\mathbf{y})$. For simple binary hypothesis

$$\frac{f(\mathbf{y}|\theta_1)}{f(\mathbf{y}|\theta_0)} \underset{<}{>} \frac{\pi_0(C_{10} - C_{00})}{\pi_1(C_{01} - C_{11})}$$

Pros and Cons:

- The knowledge of $\pi(\theta)$ allows a simple tradeoff among risks $R_{\boldsymbol{\theta}}(\delta)$.
- Bayesian detectors are widely used in digital communications where one can design the prior of the hypotheses (messages).
- Due to their simplicity, Bayesian detectors often serve as a mathematical device to obtain other results when the prior is unknown.
- The prior may not be available in practice, which limits the application of Bayesian detectors.
- If the prior is not accurate, the performance is not robust.

Notations

Problem Formulation

Λ, Λ_i	Parameter space. $\Lambda = \bigcup \Lambda_i$
$\pi(\boldsymbol{\theta})$, π_i	Prior probability density or mass function. $\pi_i = \Pr(\Theta \in \Lambda_i)$
$\mathcal{H}_i$	Hypothesis i.
$f(\mathbf{y}\|\theta)$	Conditional PDF/PMF. Likelihood function.
$f(y\|\mathcal{H}_i)$, $f_i(y)$	Conditional PDF/PMF. Likelihood function under $\mathcal{H}_i$
$L(\mathbf{y};\theta_i,\theta_j)$	Likelihood ratio $L(\mathbf{y};\theta_i,\theta_j) = \frac{f(\mathbf{y}\|\theta_i)}{f(\mathbf{y}\|\theta_j)}$
$l(\mathbf{y};\theta_i,\theta_j)$	Log-likelihood ratio $l(\mathbf{y};\theta_i,\theta_j) = \log L(\mathbf{y};\theta_i,\theta_j)$
Γ	Observation space.
Γ_i	Decision region for $\mathcal{H}_i$.

Detector

$\delta(\mathbf{y})$	Decision function (probability vector) that maps $\mathbf{y}$ to $(\Pr(D = i\|\mathbf{Y} = \mathbf{y}))$. For deterministic detectors, $\delta($$\mathbf{y}$ maps $\mathbf{y}$ directly to $\mathcal{H}_i$.
$\delta_{B,\pi}$	the Bayesian detector for prior π.

Costs and Risks

$C(i,\theta)$	Cost of detection i associated with parameter θ.
C_{ij}	Uniform cost: $C_{ij} = C(i,\theta), \theta \in \Lambda_j$.
$r(\mathbf{y})$	Likelihood ratio: $l(\mathbf{y}) = \frac{p(\mathbf{y};\theta_1)}{p(\mathbf{y};\theta_0)}$
$R_{\boldsymbol{\theta}}(\delta)$	Risk associated with parameter θ: $R_\theta(\delta) = E\{\delta(\mathbf{y})C(1,\theta) + (1-\delta(\mathbf{y}))C(0,\theta)\}$
$R(\delta)$	Bayesian risk of δ: $R(\delta) = \int p(\theta) R_\theta(\delta)$.
$R(\delta\|\mathbf{y})$	Conditional Bayesian risk: $R(\delta\|\mathbf{y}) = E\{\delta(\mathbf{y})C(1,\theta) + (1-\delta(\mathbf{y}))C(0,\theta)\|\mathbf{y}\}$
$V(\pi_0)$	The minimum bayesian risk: $V(\pi_0) = \pi_0 R_{\theta_0}(\delta_{B,\pi_0}) + (1-\pi_0) R_{\theta_1}(\delta_{B,\pi_0})$